Data mining:

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

The data sources can include database, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically.

Data warehouse:

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.

Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.

* data mining as a synonym for another popularly used term, knowledge discovery from data or KDD.

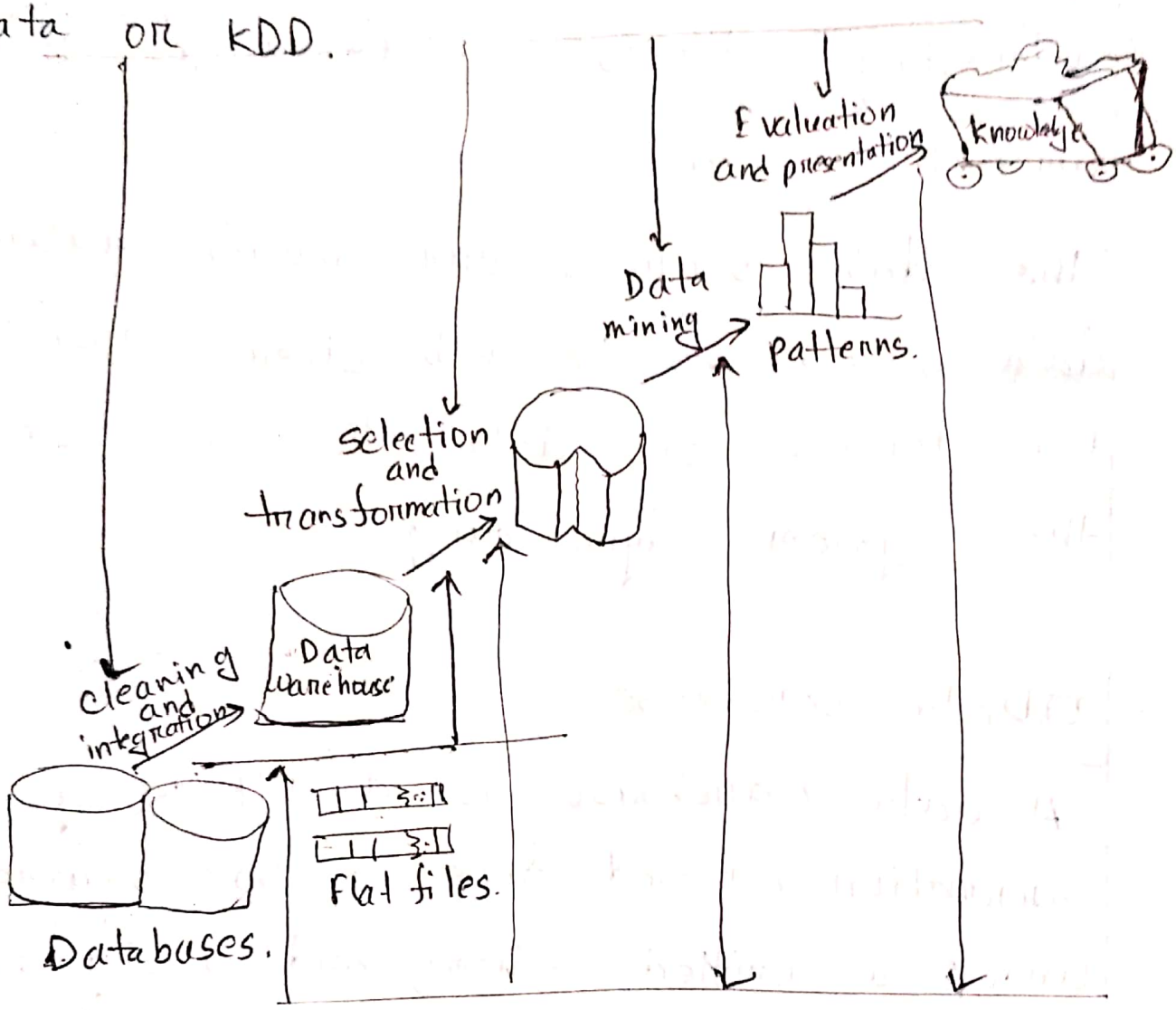


fig 1.4: Data mining as a step in the process of knowledge discovery.

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

5. Data mining (an essential process where intelligent methods are applied to extract data patterns).

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interesting measures).

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users.

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining, the data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

*for data cleaning, two types of cleanning.

① Duplicate data cleanning.

② missing data.

Redundency data:

It is a condition created within a database or data storage technology in which the same piece of data is held in two separate place.

Quantiles:

Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.

Quartiles:

It is a statistical term describing a division of observations into four defined intervals based upon the values of the data and how they compare to the entire set of observations.

two types of quantiles.

① 1st quantiles.

② 3rd "

→ The quantiles give an indication of a distribution's center, spread and shape.

→ The first quantile, denoted by Q_1, is the 25th percentile. It cuts off the lowest 25% of the data.

→ The third quantile, denoted by Q_3, is the 75th percentile. It cuts off the lowest 75% of the data.

→ The second quantile is the 50th percentile, As the median, it gives the center of the data distribution.

→ The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (IQR) and is defined as.

$$IQR = Q_3 - Q_1.$$

→ A ~~eco~~ common rule of thumb for identifying suspected outliers is to single out values falling at least $1.5 \times IQR$ above the third quartile or below the first quartile.

Boxplot:

Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:

(i) Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.

(ii) The median is marked by a line within the box.

(iii) Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

Exercise 2.2

Suppose 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 24, 29, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

(a) find 1st quantile and 3rd quantile.

(b) show boxplot of the data.

Solution:

first quantile, $Q_1 = 0.25 \times 27$ th position

$= 6.75$ " "

$= \frac{20+20}{2}$

$= 20$

Third " , $Q_3 = 0.75 \times 27$ " "

$= 20.25$ " "

$= \frac{35+35}{2}$

$= 35$

$\therefore IQR = (35 - 20) = 15$

$1.5 \times IQR = 1.5 \times 15 = 22.5$

$\therefore$ upper outlaien limit : $Q_3 + 22.5 = 57.5$

$\therefore$ lower " " : $Q_1 - 22.5 = -2.5$

$= 13$

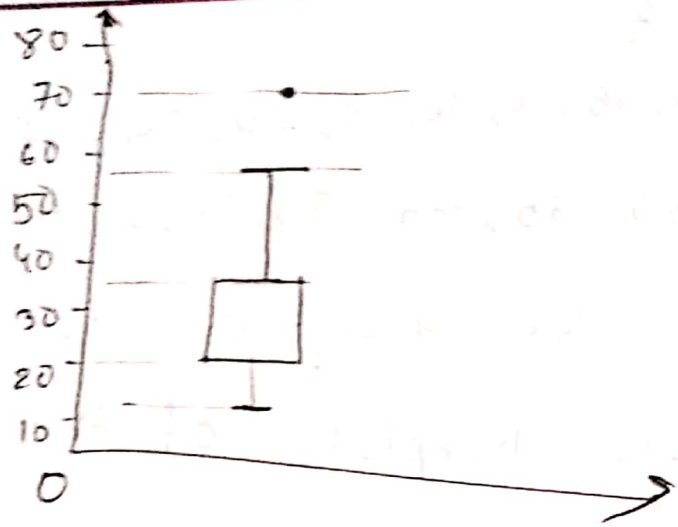


fig: Boxplot for the age values of the data.

Chap 03

II Data Reduction:

Data Reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

Data Reduction Strategies -> 3.4.1 (99page)

3.5.1 data transformation:

In data transformation, the data are transformed or consolidated into forms appropriate for mining.

Normalizations!

Normalizing the data attempts to give all attributes an equal weight. Normalization is particularly useful for classification algorithms involving neural networks or distance measurements such as nearest-neighbor classification and clustering.

There are many methods for data normalization.

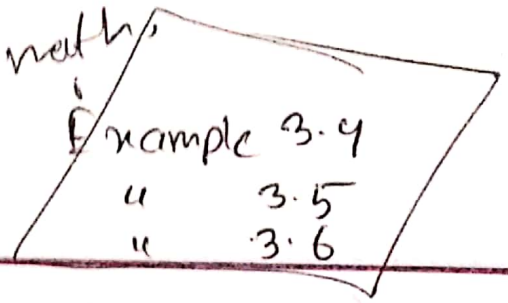
(i) Min-max normalization.

(ii) z-score "

(iii) decimal scaling.

Min-Max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute. A min-max normalization maps a value, v_i, of A to v_i' in the range $[new\_min_A, new\_max_A]$ by computing.

$$v_i' = \frac{v_i - minA}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min$$



Example 3.4 min-max normalization.